Raghubar C., Shahbazi N., Bortoluzzi, B., An, A. and McArthur J. (2017). "Integrating Visual Analytics and Machine Learning into BIM-Enabled Facilities Management". In: *LC³ 2017: Volume I – Proceedings of the Joint Conference on Computing in Construction (JC³)*, July 4-7, 2017, Heraklion, Greece, pp. 113-120. DOI: https://doi.org/10.24928/JC3-2017/0143.

INTEGRATING VISUAL ANALYTICS AND MACHINE LEARNING INTO BIM-ENABLED FACILITIES MANAGEMENT

Christopher Raghubar[1], Nima Shahbazi[2], Brandon Bortoluzzi[3], Aijun An[4], and J.J. McArthur[5]

Abstract: Building Information Modelling is becoming increasingly used for Asset Information Management in Facility Operations, where semantic and relational information are of primary importance. "Big Data" analytics tools provide new opportunities within this domain to classify and synthesize data, integrate it with the Computer-Aided Facilities Management system, and develop predictive models to assign priority and resources to address issues arising. The resulting information integrated into building information models provides a powerful tool for facilities management teams to prioritize and streamline operations and maintenance tasks.

This paper presents the development, comparison, and application of two supervised machine learning models to classify and evaluate maintenance requests generated both from within the maintenance team and occupant complaints. Three algorithms: Term Frequency (TF), Term Frequency-Inverse Category Frequency (TF-ICF), and Random Forest are used to analyse the text of the maintenance request description and assign problem types to each. Approximately 150,000 historical maintenance requests were used for model development and the models have overall prediction accuracies of 69%, 70%, and 90% for problem type prediction, respectively.

Keywords: machine learning, building information modelling, visual analytics, facility management, predictive models, big data.

1 Background

Facilities Management (FM) activities generate significant quantities of building data and information and there has been significant research in recent years to integrate it into the Building Information Modelling (BIM) environment (Volk, et al., 2014; Ilter & Ergen, 2015). BIM has a relatively low adoption rate in this context, particularly when compared to implementation in the design and construction phases (Bryde, et al., 2013) (Eadie, et al., 2013), The majority of BIM-FM research has focused on improving the geometric accuracy of BIM models, however some studies focus on relationship and trend

[1] M.A.Sc. Candidate, Department of Architectural Science, Ryerson University, Toronto, Canada christopher.raghubar@ryerson.ca

[2] PhD Candidate, Department of Electrical Engineering and Computer Science York University, Toronto, Canada, nima@cse.yorku.ca

[3] M.Arch Candidate, Department of Architectural Science, Ryerson University, Toronto, Canada brandon.bortoluzzi@ryerson.ca

[4] Professor, Department of Electrical Engineering and Computer Science York University, Toronto, Canada, aan@cse.yorku.ca

[5] Assistant Professor, Department of Architectural Science, Ryerson University, Toronto, Canada, jjmcarthur@ryerson.ca

identification to support root cause analysis, allowing for FM teams to prioritize and nullify maintenance requests based on spatial and temporal patterns. One study (Motamedi, et al., 2014) considered both the spatial and logical (systems) relationships between equipment failures to analyse underlying causes, while another (Akcamete, et al., 2010) maps maintenance and repair maintenance requests to inform root cause analysis of such failures.

A case study at Ryerson University is underway in collaboration with the Campus Facilities department to investigate how data analytics and BIM can be used to reduce the time to resolution of maintenance requests. Two problems are addressed in this case study. First, the collection and classification of such requests is often a bottleneck and a means to automatically classify such requests provides value. Second, the visualization of clusters - both spatially and temporally - is necessary to identify common problem areas and seasonal trends to inform root cause analysis of maintenance issues. Machine learning has the potential to address the former, while BIM visualization using the outputs of the machine learning addresses the latter issue. Solutions developed within both domains are thus presented herein. This forms the basis for a long-term research project to not only classify and visualize maintenance requests, but also develop a recommender system to collect additional information necessary to facilitate root cause identification and analysis, assign priority and urgency levels, and prioritize FM response to requests.

This research contributes to current discourse evaluating how BIM can be leveraged throughout the building lifecycle and how Big Data analytics tools can be applied to further enhance BIM use in the Facilities Management context.

2 METHODOLOGY

The development of the data mining and machine learning algorithms and the BIM visualization of the resulting information was undertaken in four steps: data collection and verification, data mining, machine learning algorithm development, and integration of this data with the BIM models.

2.1 Data Collection and Verification

The data set used in this study was a collection of 146,252 maintenance requests submitted to the Campus Facilities department at Ryerson University between January 2010 and December 2015. These requests were submitted through three means, listed from most to least complete information typically provided: (1) reporting through an online form with prescribed fields; (2) telephone calls to the help desk; and (3) emails sent to the maintenance team through a designated email address, where information collected was limited to the occupant-provided description of the problem, along with their email address and email timestamp. All requests are analysed by help desk staff to assign a *problem type category* (e.g. HVAC) and sub-category (e.g. "Too Cold"), before the supervisor issues a work order based on this request to the relevant trade(s).

Because supervised learning algorithms were used, ensuring accuracy of the data set was imperative to avoid training the models using incorrect labels (problem type categories and subcategories). Mislabelled data would result in incorrect classification and decision boundaries in the predictive model, thus it was necessary to review the quality of the data available. A review of 45,000 maintenance requests found an error rate less than 1% for data labelling - corrected during this review process - and remaining dataset was deemed acceptable for use in initial model training.

2.2 Data Mining

Data mining is the process to extract usable information from large quantities of data. Typical activities include counting and classifying data entries frequency, date, keyword, and/or location. The Map-Reduce algorithm (Dean & Ghemawat, 2008) is a commonly-used approach for the sorting and classification of large data sets, and was used to sort keywords associated with each problem type. This algorithm was used to mine the following data trends in work requests: (1) total frequency by month and year, (2) frequency by problem type, (3) frequency by building, and combinations thereof.

For integration within the BIM model, building, level and room parameters were created to store and display maintenance requests by period (currently open, rolling 30-day window, rolling 6-month window, rolling yearly window) as well as room-specific complaints for thermal, noise, and leak-related issues, which were determined most likely to identify underlying equipment or system problems.

2.3 Machine Learning

Machine learning approaches are broadly classified into two categories: supervised, where correct data labels are available and the machine learning algorithm is able to develop sorting hypotheses based on the training data labels, test the hypothesis for predicting labels for a test set, and compare these with the actual labels; and unsupervised learning, where no labels are available for training.

Data consistent to all entries included the location of the problem, problem description, and the contact information of the request submitter. Based on this data, the help desk manually labelled this data with problem type and subtype. The historical labelled data was thus available to support a supervised machine learning approach. One year of data (2015) was set aside as a test set and further divided into two files: an unlabelled data set and test labels. The remaining years were analysed as training data.

Data pre-processing preceded machine learning to remove unnecessary words and improve accuracy. Next, a stemming algorithm was used to obtain root words.

Three learning methods were used to learn representative words for each problem type category and classifiers for problem descriptions. In the Term Frequency (TF) method, the set of work descriptions for each problem type was analysed and the most frequently used words ("seed words") for each category were determined and used to represent that category. Each unlabelled maintenance request was then scored against each category based on the number of seed words from that category present and assigned to the category with the highest score. In the Term Frequency-Inverse Category Frequency (TF-ICF) model, the prevalence of seed words within the work description were weighted in inverse proportion to overall category frequency -- i.e. those words occurring primarily in a single problem type category were weighted more heavily than those occurring across all work description categories. In this method, the weight of the frequency term i in problem type category j is weighted using the log of the ratio of the total number of problem type categories over the number of categories in which term i appears. The resultant seed sets were thus used to assign problem types to the unlabelled test data. Finally, Random Forest (Breiman, 2001) was used, which develops multiple decision trees from the training data, creating a "forest". Each tree is grown as follows: (1) if the number of cases in the training set is N, sample N cases at random - but with replacement, from the original training data. This sample will be the training set for growing the tree; (2) if there are M features, at each node n features (where n<M) are selected at random out of the M and the best features among these m features is used to

split the node. Each tree is grown to the largest extent possible with no pruning. New instances are classified by each tree and a class probability is assigned for each. These probabilities are then used to develop an overall prediction. The Random Forest classifier is built based on the seed words extracted from the TF-ICF method, which extracted 132 features (distinct seed words) from work descriptions. Dimensionality reduction was necessary to reduce computational cost; to do so, a 10% sample of the dataset was used with the full complement of seed words to identify those most significant. The top 20 seed words were used to build the random forest classifier on the full dataset. A total of 1000 decision trees, each with a maximum depth of 8 and the Gini index to measure split quality for feature selection, were used.

2.4 BIM Integration

A virtual campus model of Ryerson University has been under development since 2014 and links computer-aided facilities management (CAFM) data to simplified geometric models of all campus buildings, which are used for the BIM visualization.

3 Evaluation

3.1 Data Mining

Several trends were noted from the data mining procedure, notably the seasonal pattern of work requests, both in terms of quantity and type (Figure 1) and in the clustering of specific work request types by building (Figure 2). The Pareto principal is illustrated in this figure where 75% of maintenance requests are being caused by 21% of the buildings on campus. Some building relationships are evident in this figure - for example, POD, JOR, and LIB make up one building and KHE, KHW, KHN and KHS are quadrants of a second, each with shared mechanical and electrical systems that trend together. Problems in one of these buildings are often symptoms of a systemic problem, and these buildings will be used in the next stage of research to develop root cause analysis strategies for maintenance request relationships building upon this research.

3.2 Machine Learning Algorithm

The first two methods used in the machine learning model produced similar, albeit different seed word fragments for each of the five most common problem types, while the random forest technique applies bootstrap aggregating to decision trees in order to obtain an average prediction of all decision trees to extract important features to use in future predictions.

Through comparison of the machine-assigned problem types to actual problem types, it was found that both methods used for seed word extraction produced reasonable overall accuracy in problem type detection: 69% for TF, 70% for TF-ICF, and 90% using problem types with the random forest model. Figure 3 presents the confusion matrices for each algorithm while Table 1 summarizes their performance, showing significant improvement of the Random Forest algorithm over both TF and TF-ICF. All models had the highest prediction accuracy in the most specific categories; this was notable particularly for the TF and TF-ICF methods. The Random Forest model significantly outperformed the other models in all categories, particularly General Maintenance, and had a weighted prediction accuracy of 95%.

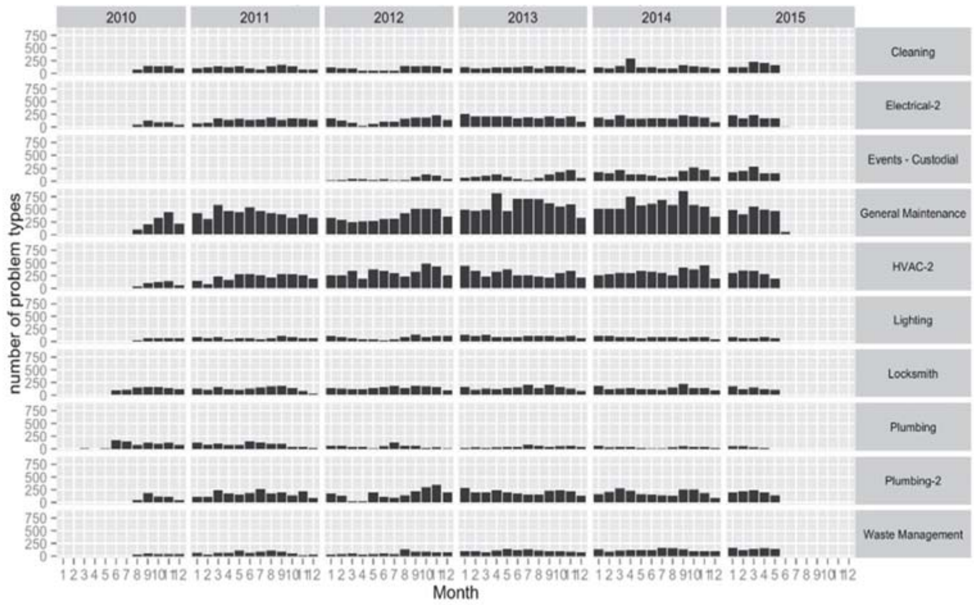


Figure 1 - Monthly Trend of Problem Type Frequency by Year

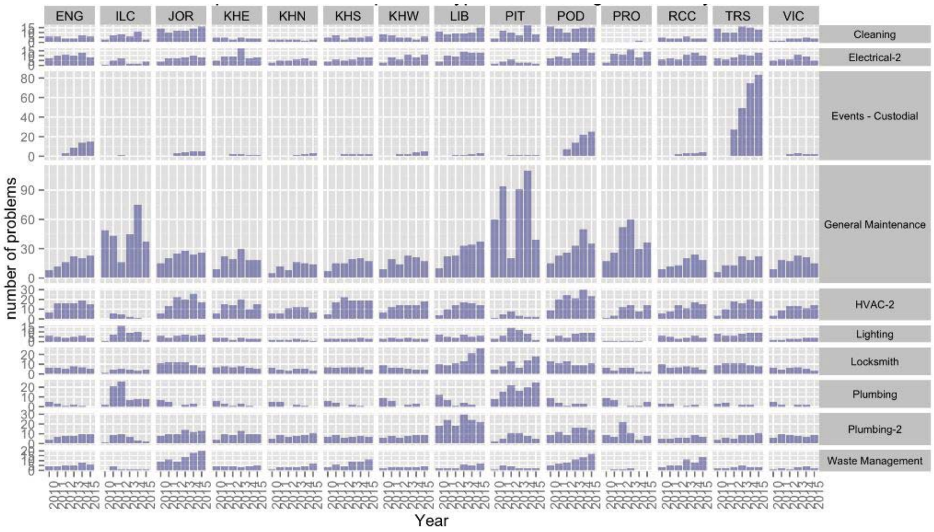


Figure 2 - Annual Work Requests by Problem Type and Building Code

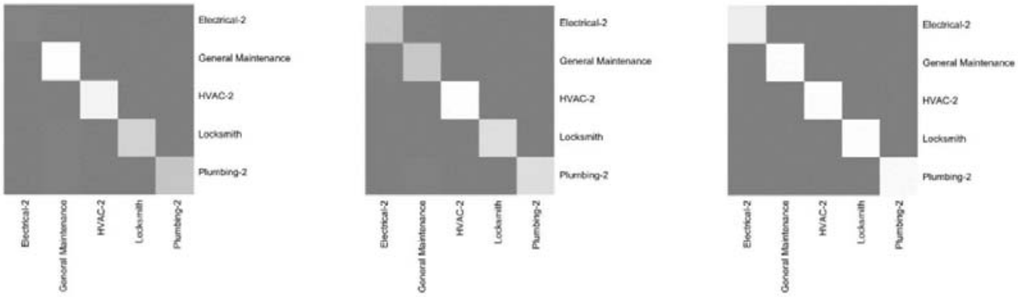


Figure 3 - Confusion Matrices for TF (left), TF-ICF (centre), and Random Forest (right)

Table 1 - Accuracy of Problem Type Prediction Based on Extraction Method

Method	Parameter	Elec.	Gen. Maint.	HVAC	Locksmith	Plumbing
TF	Precision	0.57	0.55	0.79	0.68	0.61
	Detection Rate	0.07	0.24	0.14	0.08	0.09
	Detection Prevalence	0.12	0.44	0.18	0.11	0.15
	Balanced Accuracy	0.69	0.66	0.83	0.80	0.75
TF-ICF	Precision	0.39	0.52	0.78	0.72	0.59
	Detection Rate	0.09	0.17	0.15	0.08	0.09
	Detection Prevalence	0.23	0.32	0.19	0.11	0.16
	Balanced Accuracy	0.70	0.61	0.85	0.79	0.76
Random Forest	Precision	0.96	0.84	0.99	0.91	0.91
	Detection Rate	0.11	0.38	0.19	0.1	0.12
	Detection Prevalence	0.11	0.45	0.19	0.11	0.13
	Balanced Accuracy	0.91	0.94	0.96	0.97	0.95

These results were presented to the Facilities Engineer at Ryerson University in conjunction with a procedural change proposal to implement a recommender system based on predicted problem type and cause code to solicit additional information during the work order (online) reporting process through questions targeted to identify root causes. This was seen as highly beneficial, as the quality of information obtained from occupants is inadequate in 90% of cases to determine an underlying cause, resulting in multiple trips to resolve each problem. This revised procedure is expected to streamline identification of root causes and a case study is planned for summer 2017 to test this revised procedure on 50% of the campus to quantify the impact on work order response time and personnel cost.

4 BIM VISUALIZATION STRATEGY

One of the strengths of BIM is the ability to store, synthesize, and visualize both semantic and relational information as well as geometric. The maintenance requests were integrated into the BIM model through the use of shared parameters for each problem type totalized over the desired capture window(s), in this case, both a rolling 30-day window and 5-year historical period. These parameters are instance parameters applied to each room, allowing display filters to show the relative frequency of problem types at the room, floorplan, or building level. Finer granularity visualization is possible where specific problem sub-types are of interest to the facilities management team. In this case study, one of the key questions was how many of the maintenance issues were related to thermal complaints. The Facilities Engineer wished to understand the clustering of these complaints, particularly a) whether they were higher than usual during the shoulder seasons (during heating-cooling switchover), and whether there were any zones with atypically high comfort issues. Figure 4 shows a sample floorplan with HVAC-2 | Too Cold category requests quantified in a month of particular interest.

Not visible in this figure are linked files providing access to the complete maintenance issue log at the floorplan level.

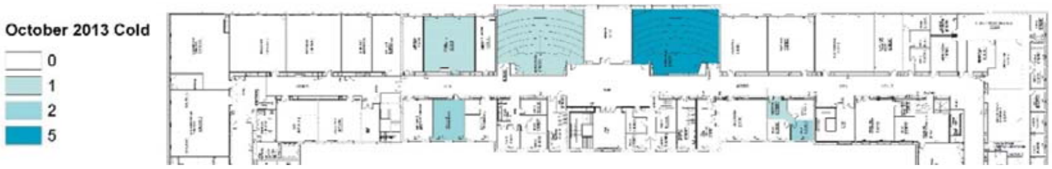
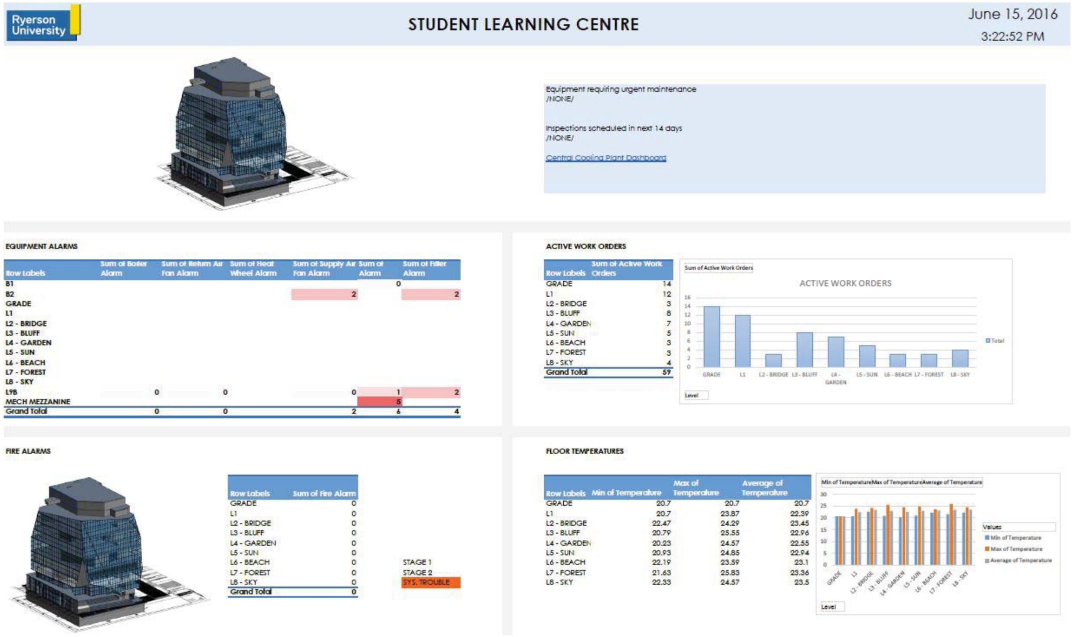


Figure 4 - Floorplan showing visualization of HVAC2 | Too Cold issues

To better consolidate this information across the campus, a dashboard was developed using Pivot Tables that read the BIM information using Dynamo. This reflects live building information and summarizes current building CAFM data. This dashboard includes a full building view, summarizing issues by floor (Fig. 5), while subsequent tabs contain floorplan views similar to that shown in Fig. 4. A time slider is being incorporated into the dashboard to allow the FM team to scroll through historical data, updating these floorplans with the parameter value associated with the selected time, and thus providing a visualization of the spatiotemporal clustering of requests of each type, to identify both priority areas and seasonal patterns of complaint.



STUDENT LEARNING CENTRE

June 15, 2016
3:22:52 PM

Equipment requiring urgent maintenance
/NONE/

Inspections scheduled in next 14 days
/NONE/

Central Cooling Plant Dashboard

EQUIPMENT ALARMS

Row Labels	Sum of Boiler Alarm	Sum of Return Air Fan Alarm	Sum of Heat Wheel Alarm	Sum of Supply Air Fan Alarm	Sum of Alarm	Sum of Filter Alarm
B1					0	
B2				2		2
GRADE						
L1						
L2 - BRIDGE						
L3 - BLUFF						
L4 - GARDEN						
L5 - SUN						
L6 - BEACH						
L7 - FOREST						
L8 - SKY						
L9B	0	0		0	1	2
MECH MEZZANINE					5	
Grand Total	0	0		2	6	4

ACTIVE WORK ORDERS

Row Labels	Sum of Active Work Orders
GRADE	14
L1	12
L2 - BRIDGE	3
L3 - BLUFF	8
L4 - GARDEN	7
L5 - SUN	5
L6 - BEACH	3
L7 - FOREST	3
L8 - SKY	4
Grand Total	59

FIRE ALARMS

Row Labels	Sum of Fire Alarm
GRADE	0
L1	0
L2 - BRIDGE	0
L3 - BLUFF	0
L4 - GARDEN	0
L5 - SUN	0
L6 - BEACH	0
L7 - FOREST	0
L8 - SKY	0
Grand Total	0

STAGE 1
STAGE 2
SYS. TROUBLE

FLOOR TEMPERATURES

Row Labels	Min of Temperature	Max of Temperature	Average of Temperature
GRADE	20.7	20.7	20.7
L1	20.7	23.87	22.39
L2 - BRIDGE	22.47	24.29	23.45
L3 - BLUFF	20.79	25.55	22.96
L4 - GARDEN	20.23	24.57	22.55
L5 - SUN	20.93	24.85	22.94
L6 - BEACH	22.19	23.59	23.1
L7 - FOREST	21.63	25.83	23.36
L8 - SKY	22.33	24.57	23.5

Figure 5 - Dashboard for sample building (full building view)

5 Conclusions

The case study presented above determined that machine learning algorithms can be used to classify highly variable occupant-generated work requests and assign problem types with a high degree of accuracy. Combined with data mining to identify trends in the CAFM data and the visualization of those trends in BIM, this provides maintenance supervisors with improved information regarding patterns of complaints, providing insight into system performance issues and other root causes. The algorithm developed in this research incorporated with a recommender system further provides the ability to tailor information collection during work order request placement, thus improving the usability of information obtained through this process.

The key findings of the machine learning algorithm development are that the accuracy of prediction is highest for the most specific problem type (subtype) categories, and that the random forest provided the highest overall accuracy, exceeding 90%. The accuracy of the predictive models could thus be further improved in future research through the following: (1) elimination/refinement of General Maintenance category in the training and testing sets and reassignment to more specific categories; (2) elimination/refinement of all *general* and - *Miscellaneous* maintenance requests for training and testing using problem subtypes, and (3) refinement of the algorithms with enhanced work requests based on recommender-system prompted information

Three limitations have also been identified for this research. First, the study of a single campus limits the application of results and replication of this research on other facility datasets is necessary for generalization. Second, the number of parameters necessary to visualize the synthesized data in BIM is onerous, thus the prioritization of problem types and sub-types is a valuable topic of future research to determine which information is of most value in facilities management and can reduce this burden. Finally, the accuracy of these models must be improved prior to real-world deployment.

The next stage of research will implement the case study discussed previously and will incorporate additional analysis to assign priority based on a combination of problem type, specific trigger words, and density of issues (geographic or temporal).

6 Acknowledgments

Funding for this research has been generously provided by the Ontario Research Fund - Research Excellence Program through the Big Data Research and Analytics Information Network (BRAIN) and FuseForward Solution Group.

7 References

Akcamete, A., Akinci, B. & Garrett, J. H., 2010. *Potential utilization of building information models for planning maintenance activities*. s.l., s.n.

Breiman, L., 2001. Random Forests. Machine Learning, 45(1), pp. 5-32.

Bryde, D., Broquetas, M. & Volm, J. M., 2013. The project benefits of building information modelling (BIM). *International Journal of Project Management*, 31(7), 971-980.

Dean, J. & Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters.. *Communications of the ACM*, 51(1), 107-113.

Eadie, R. et al., 2013. BIM implementation throughout the UK construction project lifecycle: an analysis. *Automation in Construction*, 36, 145-151.

Ilter, D. & Ergen, E., 2015. BIM for building refurbishment and maintenance: current status and research directions.. *Structural Survey*, 33(3), 228-256..

Motamedi, A., Hammad, S. & Asen, Y., 2014. Knowledge-assisted BIM-based visual analytics for failure root cause detection in facilities management.. *Automation in Construction*, 43, 73-83.

Volk, R., Stengel, J. & Scultmann, F., 2014. Building Information Modeling (BIM) for existing buildings — Literature review and future needs. *Automation in Construction*, 38, 109-127.